Оптимизация вычислений нейронных сетей на графических процессорах для мобильных устройств

Оплачко Николай Алексеевич

Кафедра системного программирования

e-mail: oplachko@ispras.ru

Научный руководитель — д.ф.-м.н. проф. ак. РАН Аветисян Арутюн Ишханович

В последние годы нейронные сети широко используются для решения задач в различных областях. Иногда возникает потребность запускать нейронные сети на мобильных устройствах в силу разных причин, в частности: защита данных пользователя, сокращение затрат на серверы, уменьшение временной задержки, обеспечение возможности использования приложения без доступа к сети Интернет.

Вследствие ограниченности ресурсов мобильных устройств критична оптимизация вычислений предсказаний нейронных сетей для сокращения расходов времени и используемой памяти.

Выпускная квалификационная работа включает в себя следующие задачи:

— исследование существующих подходов для оптимизации затраченного времени и используемой памяти при вычислении предсказаний нейронных сетей;

— разработка метода одновременной оптимизации времени и памяти;

— реализация и встраивание предложенного метода в библиотеку искусственного интеллекта с использованием прикладного программного интерфейса Vulkan.

В работе рассматриваются задача оптимизации памяти, решаемая эвристическими алгоритмами [1, 2], и задача параллельных вычислений на уровне ациклического графа вычислений, решаемая нахождением некоторого разбиения графа на группы независимых друг от друга вершин [3], далее называемых слоями.

Отмечается, что эти задачи конфликтуют между собой: параллелизация вычислений накладывает дополнительные ограничения на повторное использование памяти.

Предлагается метод, позволяющий находить компромисс в конфликте вышеупомянутых задач. Суть его состоит в разбиении найденных слоёв на несколько групп: с ростом числа групп ослабляются добавляемые ограничения на повторное использование памяти, при этом понижается степень параллелизма.

Возможность конфигурировать предложенный метод для используемой нейронной сети (варьировать количество групп при делении слоёв) позволяет разработчику выбрать компромисс между затрачиваемыми на вычисления временем и памятью.

В ходе тестирования предложенного метода для вычисления предсказания свёрточной нейронной сети ResNeXt [4] наилучший результат был получен

при разбиении слоёв на две группы: ускорение составило 6.85% при сокращении потребления памяти на 72.4%. Оказалось, что при разбиении слоёв на одну группу вместо двух, кроме того, что на 2.8% увеличивается используемая память, значимо падает ускорение: 3.08% вместо 6.85%. Это означает, что особенности обращений в память графического процессора могут давать существенный вклад в итоговое время вычислений наравне с параллелизацией.

Тестирование предложенного метода производилось на одноплатном компьютере HiKey 970. Была подсчитана электрическая энергия во время вычисления предсказания нейронной сети ResNeXt при помощи цифрового измерителя мощности. Из результатов для этой нейронной сети следует, что с ростом степени параллелизма растут суммарные затраты электрической энергии, поэтому с целью баланса экономии затрат времени, памяти и электрической энергии может оказаться оптимальным разбиение слоёв более, чем на две группы. Также было произведено профилирование графического процессора мобильного устройства набором инструментов Arm Development Studio. Подсчёт числа тактов ожидания обращений в память подтвердил, что при разбиении слоёв на одну группу работа с памятью менее оптимальна, чем на две.

Основной результат данной работы состоит в разработке метода для одновременной оптимизации времени и памяти при вычислении предсказаний нейронных сетей, его реализации и встраивании в программно-аппаратную часть библиотеки машинного обучения, а также получении положительных результатов при его тестировании.

Литература

[1] Pisarchyk Y., Lee J. Efficient memory management for deep neural net inference // arXiv preprint arXiv:2001.03288. 2020.

[2] Profile-guided memory optimization for deep neural networks / Sekiyama T., Imamichi T., Imai H., Raymond R. // arXiv preprint arXiv:1804.10001. 2018.

[3] Node-level parallelization for deep neural networks with conditional independent graph / Zhou F., Wu F., Zhang Z., Dong M. // Neurocomputing, 267. 2017. P. 261–270.

[4] Aggregated residual transformations for deep neural networks / Xie S., Girshick R., Dollár P., Tu Z., He K. // Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. P. 1492–1500.